

**REMARKS**

Claims 1-30 are pending. Claims 1, 3-6, 8-13, 15-17, 19, 21-22, and 28 stand rejected under 35 U.S.C. § 102(b) as being anticipated by U.S. Patent No. 5,619,709 to Caid et al. Claims 2, 25-27, and 29-30 stand rejected under 35 U.S.C. § 103(a) as being unpatentable over U.S. Patent No. 5,619,709 to Caid et al. in view of U.S. Patent No. 6,470,307 to Turney. Claim 7 stands rejected under U.S.C. § 103(a) as being unpatentable over U.S. Patent No. 5,619,709 to Caid et al. in view of U.S. Patent No. 5,778,397 to Kupiec et al. Claims 20 and 23-24 stand rejected under 35 U.S.C. § 103(a) as being unpatentable over U.S. Patent No. 5,619,709 to Caid et al. in view of U.S. Patent No. 6,289,353 to Hazlehurst et al. Claim 14 stands rejected under 35 U.S.C. § 103(a) as being unpatentable over U.S. Patent No. 5,619,709 to Caid et al. in view of U.S. Patent No. 6,173,261 to Arai et al. Claim 18 stands rejected under 35 U.S.C. § 103(a) as being unpatentable over U.S. Patent No. 5,539,841 to Huttenlocher et al.

Reconsideration is requested. No new matter is added. The rejections are traversed. Claims 1, 5, 11, 17, 23, 25, and 29 are amended. Claim 28 is canceled. Claims 31-41 are added. Claims 1-27 and 29-41 remain in the case for consideration.

**REJECTIONS UNDER 35 U.S.C. § 102(b)**

Referring to claim 1, the invention is directed toward a method for determining dominant phrase vectors in a topological vector space for a semantic content of a document on a computer system, the method comprising: identifying a directed set of concepts as a dictionary, the directed set including a maximal element at least one concept, and at least one chain from the maximal element to every concept; selecting a subset of the chains to form a basis for the dictionary; accessing dominant phrases for the document, the dominant phrases representing a condensed content for the document; measuring how concretely each dominant phrase is represented in each chain in the basis and the dictionary; constructing at least one state vector in the topological vector space for each dominant phrase using the measures of how concretely each dominant phrase is represented in each chain in the dictionary and the basis; and collecting the state vectors into the dominant phrase vectors for the document.

Referring to claim 5, the invention is directed toward a method for determining dominant vectors in a topological vector space for a semantic content of a document on a computer system, the method comprising: identifying a directed set of concepts as a dictionary, the directed set including a maximal element at least one concept, and at least one chain from the maximal element to every concept; selecting a subset of the chains to form a

basis for the dictionary; storing the document in computer memory accessible by the computer system; extracting words from at least a portion of the document; measuring how concretely each word is represented in each chain in the basis and the dictionary; constructing a state vector in the topological vector space for each word using the measures of how concretely each word is represented in each chain in the dictionary and the basis; filtering the state vectors; and collecting the filtered state vectors into the dominant vectors for the document.

Referring to claim 11, the invention is directed toward a method for determining a semantic abstract in a topological vector space for a semantic content of a document on a computer system, the method comprising: identifying a directed set of concepts as a dictionary, the directed set including a maximal element at least one concept, and at least one chain from the maximal element to every concept; selecting a subset of the chains to form a basis for the dictionary; storing the document in computer memory accessible by the computer system; determining dominant phrases for the document; measuring how concretely each dominant phrase is represented in each chain in the basis and the dictionary; constructing dominant phrase vectors in the topological vector space for the dominant phrases using the measures of how concretely each dominant phrase is represented in each chain in the dictionary and the basis; selecting words for the document; measuring how concretely each word is represented in each chain in the basis and the dictionary; constructing dominant vectors in the topological vector space for the words using the measures of how concretely each word is represented in each chain in the dictionary and the basis; and generating the semantic abstract using the dominant phrase vectors and the dominant vectors.

Referring to claim 17, the invention is directed toward method for comparing the semantic content of first and second documents on a computer system, the method comprising: identifying a directed set of concepts as a dictionary, the directed set including a maximal element at least one concept, and at least one chain from the maximal element to every concept; selecting a subset of the chains to form a basis for the dictionary; accessing dominant phrases for the first document, the dominant phrases representing a condensed content for the first document; measuring how concretely each dominant phrase for the first document is represented in each chain in the basis and the dictionary; constructing at least one state vector for the first document in the topological vector space for each dominant phrase for the first document using the measures of how concretely each dominant phrase for the first document is represented in each chain in the dictionary and the basis; collecting the state vectors for the first document into the semantic abstract for the first document;

determining a semantic abstract for the second document; measuring a distance between the semantic abstracts; and classifying how closely related the first and second documents are using the distance.

Referring to claim 23, the invention is directed toward a method for locating a second document on a computer with a semantic content similar to a first document, the method comprising: identifying a directed set of concepts as a dictionary, the directed set including a maximal element at least one concept, and at least one chain from the maximal element to every concept; selecting a subset of the chains to form a basis for the dictionary; accessing dominant phrases for the first document, the dominant phrases representing a condensed content for the first document; measuring how concretely each dominant phrase for the first document is represented in each chain in the basis and the dictionary; constructing at least one state vector for the first document in the topological vector space for each dominant phrase for the first document using the measures of how concretely each dominant phrase for the first document is represented in each chain in the dictionary and the basis; collecting the state vectors for the first document into the semantic abstract for the first document; locating a second document; determining a semantic abstract for the second document; measuring a distance between the semantic abstracts for the first and second documents; classifying how closely related the first and second documents are using the distance; and if the second document is classified as having a semantic content similar to the semantic content of the first document, selecting the second document.

Referring to claim 25, the invention is directed toward an apparatus on a computer system to determine a semantic abstract in a topological vector space for a semantic content of a document stored on the computer system, the apparatus comprising: a phrase extractor adapted to extract phrases from the document; a state vector constructor adapted to construct at least one state vector in the topological vector space for each phrase extracted by the phrase extractor, the state vectors measuring how concretely each phrase extracted by the phrase extractor is represented in each chain in a basis and a dictionary, the dictionary including a directed set of concepts including a maximal element and at least one chain from the maximal element to every concept in the directed set, the basis including a subset of chains in the directed set; and collection means for collecting the state vectors into the semantic abstract for the document.

Referring to claim 29, the invention is directed toward a method for determining a semantic abstract in a topological vector space for a semantic content of a document on a computer system, the method comprising: extracting dominant phrases from the document

using a phrase extractor, the dominant phrases representing a condensed content for the document; identifying a directed set of concepts as a dictionary, the directed set including a maximal element at least one concept, and at least one chain from the maximal element to every concept; selecting a subset of the chains to form a basis for the dictionary; measuring how concretely each dominant phrase is represented in each chain in the basis and the dictionary; constructing at least one first state vector in the topological vector space for each dominant phrase using the measures of how concretely each dominant phrase is represented in each chain in the dictionary and the basis; collecting the first state vectors into dominant phrase vectors for the document; extracting words from at least a portion of the document; constructing a second state vector in the topological vector space for each word using the dictionary and the basis; filtering the second state vectors; collecting the filtered second state vectors into dominant vectors for the document; and generating the semantic abstract using the dominant phrase vectors and the dominant vectors.

In contrast, Caid teaches a system and method for generating context vectors. Caid begins by selecting the dimension for the context vectors: Caid recommends 200 or more components, and this number is fixed in advance. Caid then generates random numbers to initialize the components of the context vectors: in other words, the context vectors are generated randomly.

Caid then works through the words in the documents one by one. Applying the appropriate learning laws, Caid adjusts the context vectors to account for the "importance" of a word and its neighbor. In this way, Caid produces vectors that, in the end, reflect the context of the words the vectors represent. The document can then be represented by a summary vector, which somehow combines the significance of the context vectors for the individual words.

The problem with Caid as a reference stems from the fact that Caid uses a particular technique to generate the context vectors; this technique differs greatly from that used in the claimed invention. As discussed above, Caid starts by generating random context vectors, expecting that the dot product of any pair will be close to zero (signifying that there is no relationship between the words represented by the context vectors). The context vectors are then changed slowly as the documents are analyzed, to adjust the context vectors to their final values.

In rejecting claim 1, the Examiner has argued that Caid teaches constructing state vectors using a dictionary and a basis. The problem is that nowhere does Caid teach anything analogous to the concept of a "basis" in the claimed invention. A "basis" is a set of chains in

the directed set. Each concept can then be measured relative to each basis chain to determine how concretely the concept is represented in each basis chain; these measurements can then be used to form the state vectors for the concepts. The Examiner is referred to parent U.S. Patent Application Serial No. 09/512,963, which has been incorporated by reference into the present application, for more information, specifically with respect to pages 11-18, wherein the concepts of chains, bases, and how concepts can be measured relative to basis chains are all discussed.

It is worth noting that Caid generates the context vectors in a very different manner than the claimed invention. In the claimed invention, the state vectors are generated by measuring a concept against a basis set of chains in the directed set. The concepts in the directed set are typically determined and organized before a document is analyzed. While new concepts can be added, doing so can require changing the state vectors for a concept. In Caid, the context vectors are initially random, and are slowly "walked" to their final value, as documents are examined.

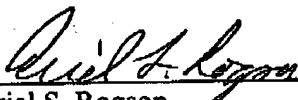
To aid the Examiner in his interpretation of the claims, claims 1, 5, 11, 17, 23, 25, and 29 have been amended to include details of how the state vectors are constructed. Because Caid teaches a very different system for generating context vectors, Caid cannot teach the features of the claims (including the concept and use of a "basis", as well as the features added to clarify the meaning of this term). Accordingly, Caid cannot anticipate the claims, and claims 1-27 and 29-41 are allowable under 35 U.S.C. § 102(b) over Caid et al.

An additional difference between Caid and the claimed invention is that in Caid it is unlikely that a given word will always have the same context vector. A close look at FIG. 3 of Caid shows that the context vectors for the first document are generated before the second document is analyzed. This suggests that the context vector for a given word in the first document might be a different vector than the vector for the same word in the last document, simply because the context vector has been refined during the document analysis. In contrast, for the invention, given a particular dictionary and basis, the state vector is constant, regardless of the number of documents analyzed with the dictionary. This concept is expressed in new claims 35-41, and is supported in the specification of parent U.S. Patent Application Serial No. 09/512,963, which has been incorporated by reference into the present application. The background is provided in the attached copies of the parent patent application. Specifically, the Examiner is asked to note the construction of the functions  $g_k(s)$ , which define the individual components of a vector for a concept in the directed set. Given a particular directed set (that is, a particular dictionary) and a selected set of basis

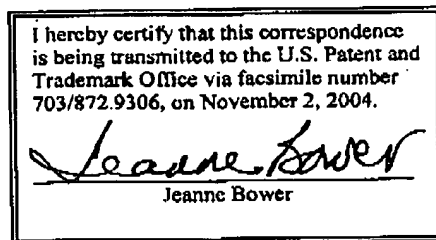
chains, the vector for a concept can be calculated directly, without reference to any documents. In other words, the documents are used only to identify the terms (dominant phrases and/or words) relating to the document, so that the pertinent vectors can be selected. It should be apparent to the Examiner that, in fact, all of the vectors in the topological vector space can be computed in advance of examining any document. For example, the example vectors shown at the top of page 17 of parent U.S. Patent Application Serial No. 09/512,963 have been computed without reference to any particular document. This shows that vector construction can be performed independent of any document; the document is pertinent only in identifying the concepts for which vectors are desired. Thus, claims 35-41 are supported by the specification of parent U.S. Patent Application Serial No. 09/512,963; and since parent U.S. Patent Application Serial No. 09/512,963, has been incorporated by reference into the present application, this patent application supports claims 35-41.

For the foregoing reasons, reconsideration and allowance of claims 1-27 and 29-41 of the application as amended is solicited. The Examiner is encouraged to telephone the undersigned at (503) 222-3613 if it appears that an interview would be helpful in advancing the case.

Respectfully submitted,  
MARGER JOHNSON & McCOLLOM, P.C.

  
Ariel S. Rogson  
Reg. No. 43,054

MARGER JOHNSON & McCOLLOM, P.C.  
1030 SW Morrison Street  
Portland, OR 97205  
503-222-3613  
Customer No. 20575



*An Example Topology*

Consider an actual topology on the set  $P$  of predicates. This is accomplished by exploiting the notion of hyponymy and meaning postulates.

Let  $P$  be the set of predicates, and let  $B$  be the set of all elements of  $2^P$ , i.e.,  $\wp(\wp(P))$ , that express hyponymy.  $B$  is a basis, if not of  $2^P$ , i.e.,  $\wp(P)$ , then at least of everything worth talking about:  $S = \cup \{b : b \in B\}$ . If  $b_\alpha, b_\gamma \in B$ , neither containing the other, have a non-empty intersection that is not already an explicit hyponym, extend the basis  $B$  with the meaning postulate  $b_\alpha \cap b_\gamma$ . For example, "dog" is contained in both "carnivore" and "mammal." So, even though the core lexicon may not include an entry equivalent to "carnivorous mammal," it is a worthy meaning postulate, and the lexicon can be extended to include the intersection. Thus,  $B$  is a basis for  $S$ .

Because hyponymy is based on nested subsets, there is a hint of partial ordering on  $S$ . A partial order would be a big step towards establishing a metric.

At this point, a concrete example of a (very restricted) lexicon is in order. FIG. 3 shows a set of concepts, including "thing" 305, "man" 310, "girl" 312, "adult human" 315, "kinetic energy" 320, and "local action" 325. "Thing" 305 is the maximal element of the set, as every other concept is a type of "thing." Some concepts, such as "man" 310 and "girl" 312 are "leaf concepts," in the sense that no other concept in the set is a type of "man" or "girl." Other concepts, such as "adult human" 315, "kinetic energy" 320, and "local action" 325 are "internal concepts," in the sense that they are types of other concepts (e.g., "local action" 325 is a type of "kinetic energy" 320) but there are other concepts that are types of these concepts (e.g., "man" 310 is a type of "adult human" 315).

FIG. 4 shows a directed set constructed from the concepts of FIG. 3. For each concept in the directed set, there is at least one chain extending from maximal element "thing" 305 to the concept. These chains are composed of directed links, such as links 405, 410, and 415, between pairs of concepts. In the directed set of FIG. 4, every chain from maximal element "thing" must pass through either "energy" 420 or "category" 425. Further, there can be more than one chain extending from maximal element "thing" 305 to any concept. For example, there are four chains extending from "thing" 305 to "adult human"

315: two go along link 410 extending out of "being" 435, and two go along link 415 extending out of "adult" 445.

Some observations about the nature of FIG. 4:

- First, the model is a *topological space*.
- Second, note that *the model is not a tree*. In fact, it is an example of a *directed set*. For example, concepts "being" 430 and "adult human" 315 are types of multiple concepts higher in the hierarchy. "Being" 430 is a type of "matter" 435 and a type of "behavior" 440; "adult human" 315 is a type of "adult" 445 and a type of "human" 450.
- Third, observe that the relationships expressed by the links are indeed relations of hyponymy.
- Fourth, note particularly – but without any loss of generality – that "man" 310 maps to both "energy" 420 and "category" 425 (via composite mappings) which in turn both map to "thing" 305; i.e., the (composite) relations are multiple valued and induce a partial ordering. These multiple mappings are natural to the meaning of things and critical to semantic characterization.
- Finally, note that "thing" 305 is *maximal*; indeed, "thing" 305 is the *greatest* element of *any* quantization of the lexical semantic field (subject to the premises of the model).

#### Metrizing S

FIGs. 5A-5G show eight different chains in the directed set that form a basis for the directed set. FIG. 5A shows chain 505, which extends to concept "man" 310 through concept "energy" 420. FIG. 5B shows chain 510 extending to concept "iguana." FIG. 5C shows another chain 515 extending to concept "man" 310 via a different path. FIGs. 5D-5G show other chains.

FIG. 13 shows a data structure for storing the directed set of FIG. 3, the chains of FIG. 4, and the basis chains of FIGs. 5A-5G. In FIG. 13, concepts array 1305 is used to store the concepts in the directed set. Concepts array 1305 stores pairs of elements. One element identifies concepts by name; the other element stores numerical identifiers 1306. For example, concept name 1307 stores the concept "dust," which is paired with numerical



identifier "2" 1308. Concepts array 1305 shows 9 pairs of elements, but there is no theoretical limit to the number of concepts in concepts array 1305. In concepts array 1305, there should be no duplicated numerical identifiers 1306. In FIG. 13, concepts array 1305 is shown sorted by numerical identifier 1306, although this is not required. When concepts array 1305 is sorted by numerical identifier 1306, numerical identifier 1306 can be called the *index* of the concept name.

Maximal element (ME) 1310 stores the index to the maximal element in the directed set. In FIG. 13, the concept index to maximal element 1310 is "6," which corresponds to concept "thing," the maximal element of the directed set of FIG. 4.

Chains array 1315 is used to store the chains of the directed set. Chains array 1315 stores pairs of elements. One element identifies the concepts in a chain by index; the other element stores a numerical identifier. For example, chain 1317 stores a chain of concept indices "6", "5", "9", "7", and "2," and is indexed by chain index "1" (1318). (Concept index 0, which does not occur in concepts array 1305, can be used in chains array 1315 to indicate the end of the chain. Additionally, although chain 1317 includes five concepts, the number of concepts in each chain can vary.) Using the indices of concepts array 1305, this chain corresponds to concepts "thing," "energy," "potential energy," "matter," and "dust." Chains array 1315 shows one complete chain and part of a second chain, but there is no theoretical limit to the number of chains stored in chain array 1315. Observe that, because maximal element 1310 stores the concept index "6," every chain in chains array 1315 should begin with concept index "6." Ordering the concepts within a chain is ultimately helpful in measuring distances between the concepts. However concept order is not required. Further, there is no required order to the chains as they are stored in chains array 1315.

Basis chains array 1320 is used to store the chains of chains array 1315 that form a basis of the directed set. Basis chains array 1320 stores chain indices into chains array 1315. Basis chains array 1320 shows four chains in the basis (chains 1, 4, 8, and 5), but there is no theoretical limit to the number of chains in the basis for the directed set.

Euclidean distance matrix 1325A stores the distances between pairs of concepts in the directed set of FIG. 4. (How distance is measured between pairs of concepts in the directed set is discussed below. But in short, the concepts in the directed set are mapped to state vectors in multi-dimensional space, where a state vector is a directed line segment starting at

the origin of the multi-dimensional space and extending to a point in the multi-dimensional space.) The distance between the end points of pairs of state vectors representing concepts is measured. The smaller the distance is between the state vectors representing the concepts, the more closely related the concepts are. Euclidean distance matrix 1325A uses the indices 1306 of the concepts array for the row and column indices of the matrix. For a given pair of row and column indices into Euclidean distance matrix 1325A, the entry at the intersection of that row and column in Euclidean distance matrix 1325A shows the distance between the concepts with the row and column concept indices, respectively. So, for example, the distance between concepts "man" and "dust" can be found at the intersection of row 1 and column 2 of Euclidean distance matrix 1325A as approximately 1.96 units. The distance between concepts "man" and "iguana" is approximately 1.67, which suggests that "man" is closer to "iguana" than "man" is to "dust." Observe that Euclidean distance matrix 1325A is symmetrical: that is, for an entry in Euclidean distance matrix 1325A with given row and column indices, the row and column indices can be swapped, and Euclidean distance matrix 1325A will yield the same value. In words, this means that the distance between two concepts is not dependent on concept order: the distance from concept "man" to concept "dust" is the same as the distance from concept "dust" to concept "man."

Angle subtended matrix 1325B is an alternative way to store the distance between pairs of concepts. Instead of measuring the distance between the state vectors representing the concepts (see below), the angle between the state vectors representing the concepts is measured. This angle will vary between 0 and 90 degrees. The narrower the angle is between the state vectors representing the concepts, the more closely related the concepts are. As with Euclidean distance matrix 1325A, angle subtended matrix 1325B uses the indices 1306 of the concepts array for the row and column indices of the matrix. For a given pair of row and column indices into angle subtended matrix 1325B, the entry at the intersection of that row and column in angle subtended matrix 1325B shows the angle subtended the state vectors for the concepts with the row and column concept indices, respectively. For example, the angle between concepts "man" and "dust" is approximately 51 degrees, whereas the angle between concepts "man" and "iguana" is approximately 42 degrees. This suggests that "man" is closer to "iguana" than "man" is to "dust." As with Euclidean distance matrix 1325A, angle subtended matrix 1325B is symmetrical.

Not shown in FIG. 13 is a data structure component for storing state vectors (discussed below). As state vectors are used in calculating the distances between pairs of concepts, if the directed set is static (i.e., concepts are not being added or removed and basis chains remain unchanged), the state vectors are not required after distances are calculated.

5 Retaining the state vectors is useful, however, when the directed set is dynamic. A person skilled in the art will recognize how to add state vectors to the data structure of FIG. 13.

Although the data structure for concepts array 1305, maximal element 1310 chains array 1315, and basis chains array 1320 in FIG. 13 are shown as arrays, a person skilled in the art will recognize that other data structures are possible. For example, concepts array 10 could store the concepts in a linked list, maximal element 1310 could use a pointer to point to the maximal element in concepts array 1305, chains array 1315 could use pointers to point to the elements in concepts array, and basis chains array 1320 could use pointers to point to chains in chains array 1315. Also, a person skilled in the art will recognize that the data in Euclidean distance matrix 1325A and angle subtended matrix 1325B can be stored using 15 other data structures. For example, a symmetric matrix can be represented using only one half the space of a full matrix if only the entries below the main diagonal are preserved and the row index is always larger than the column index. Further space can be saved by computing the values of Euclidean distance matrix 1325A and angle subtended matrix 1325B "on the fly" as distances and angles are needed.

20 Returning to FIGs. 5A-5G, how are distances and angles subtended measured? The chains shown in FIGs. 5A-5G suggest that the relation between any node of the model and the maximal element "thing" 305 can be expressed as any one of a set of *composite* functions; one function for each chain from the minimal node  $\mu$  to "thing" 305 (the  $n^{\text{th}}$  predecessor of  $\mu$  along the chain):

$$25 \quad f: \mu \Rightarrow \text{thing} = f_1 \circ f_2 \circ f_3 \circ \dots \circ f_n$$

where the chain connects  $n + 1$  concepts, and  $f_j$  links the  $(n - j)^{\text{th}}$  predecessor of  $\mu$  with the  $(n + 1 - j)^{\text{th}}$  predecessor of  $\mu$ ,  $1 \leq j \leq n$ . For example, with reference to FIG. 5A, chain 505 connects nine concepts. For chain 505,  $f_1$  is link 505A,  $f_2$  is link 505B, and so on through  $f_8$  being link 505H.

30 Consider the set of all such functions for all minimal nodes. Choose a countable subset  $\{f_k\}$  of functions from the set. For each  $f_k$  construct a function  $g_k: S \Rightarrow I^1$  as follows.

For  $s \in S$ ,  $s$  is in relation (under hyponymy) to "thing" 305. Therefore,  $s$  is in relation to at least one predecessor of  $\mu$ , the minimal element of the (unique) chain associated with  $f_k$ . Then there is a predecessor of smallest index (of  $\mu$ ), say the  $m^{\text{th}}$ , that is in relation to  $s$ . Define:

$$g_k(s) = (n - m) / n \quad \text{Equation (2)}$$

This formula gives a measure of concreteness of a concept to a given chain associated with function  $f_k$ .

As an example of the definition of  $g_k$ , consider chain 505 of FIG. 5A, for which  $n$  is 8. Consider the concept "cat" 555. The smallest predecessor of "man" 310 that is in relation to "cat" 555 is "being" 430. Since "being" 430 is the fourth predecessor of "man" 310,  $m$  is 4, and  $g_k(\text{"cat" } 555) = (8 - 4) / 8 = 1/2$ . "Iguana" 560 and "plant" 560 similarly have  $g_k$  values of  $1/2$ . But the only predecessor of "man" 310 that is in relation to "adult" 445 is "thing" 305 (which is the eighth predecessor of "man" 310), so  $m$  is 8, and  $g_k(\text{"adult" } 445) = 0$ .

Finally, define the vector valued function  $\varphi: S \Rightarrow \mathbb{R}^k$  relative to the indexed set of scalar functions  $\{g_1, g_2, g_3, \dots, g_k\}$  (where scalar functions  $\{g_1, g_2, g_3, \dots, g_k\}$  are defined according to Equation (2)) as follows:

$$\varphi(s) = \langle g_1(s), g_2(s), g_3(s), \dots, g_k(s) \rangle \quad \text{Equation (3)}$$

This state vector  $\varphi(s)$  maps a concept  $s$  in the directed set to a point in  $k$ -space ( $\mathbb{R}^k$ ). One can measure distances between the points (the state vectors) in  $k$ -space. These distances provide measures of the closeness of concepts within the directed set. The means by which distance can be measured include distance functions, such as Equations (1a), (1b), or (1c). Further, trigonometry dictates that the distance between two vectors is related to the angle subtended between the two vectors, so means that measure the angle between the state vectors also approximates the distance between the state vectors. Finally, since only the direction (and not the magnitude) of the state vectors is important, the state vectors can be normalized to the unit sphere. If the state vectors are normalized, then the angle between two state vectors is no longer an approximation of the distance between the two state vectors, but rather is an exact measure.

The functions  $g_k$  are analogous to step functions, and in the limit (of refinements of the topology) the functions are continuous. Continuous functions preserve local topology; i.e., "close things" in  $S$  map to "close things" in  $\mathbb{R}^k$ , and "far things" in  $S$  tend to map to "far things" in  $\mathbb{R}^k$ .

5

### Example Results

The following example results show state vectors  $\phi(s)$  using chain 505 as function  $g_1$ , chain 510 as function  $g_2$ , and so on through chain 540 as function  $g_8$ .

10

$$\phi(\text{"boy"}) \Rightarrow \langle 3/4, 5/7, 4/5, 3/4, 7/9, 5/6, 1, 6/7 \rangle$$

$$\phi(\text{"dust"}) \Rightarrow \langle 3/8, 3/7, 3/10, 1, 1/9, 0, 0, 0 \rangle$$

$$\phi(\text{"iguana"}) \Rightarrow \langle 1/2, 1, 1/2, 3/4, 5/9, 0, 0, 0 \rangle$$

$$\phi(\text{"woman"}) \Rightarrow \langle 7/8, 5/7, 9/10, 3/4, 8/9, 2/3, 5/7, 5/7 \rangle$$

$$\phi(\text{"man"}) \Rightarrow \langle 1, 5/7, 1, 3/4, 1, 1, 5/7, 5/7 \rangle$$

15

Using these state vectors, the distances between concepts and the angles subtended between the state vectors are as follows:

Pairs of Concepts	Distance (Euclidean)	Angle Subtended
"boy" and "dust"	~1.85	~52°
"boy" and "iguana"	~1.65	~46°
"boy" and "woman"	~0.41	~10°
"dust" and "iguana"	~0.80	~30°
"dust" and "woman"	~1.68	~48°
"iguana" and "woman"	~1.40	~39°
"man" and "woman"	~0.39	~07°

From these results, the following comparisons can be seen:

20

- "boy" is closer to "iguana" than to "dust."
- "boy" is closer to "iguana" than "woman" is to "dust."
- "boy" is much closer to "woman" than to "iguana" or "dust."
- "dust" is further from "iguana" than "boy" to "woman" or "man" to "woman."
- "woman" is closer to "iguana" than to "dust."

- “woman” is closer to “iguana” than “boy” is to “dust.”
- “man” is closer to “woman” than “boy” is to “woman.”

All other tests done to date yield similar results. The technique works consistently well.

5

### *How It (Really) Works*

As described above, construction of the  $\phi$  transform is (very nearly) an algorithm. In effect, this describes a *recipe* for metrizing a lexicon – or for that matter, metrizing anything that can be modeled as a directed set – but does not address the issue of *why* it works. In other words, *what's really going on here?* To answer this question, one must look to the underlying mathematical principles.

10

First of all, what is the nature of  $S$ ? Earlier, it was suggested that a propositional model of the lexicon has found favor with many linguists. For example, the lexical element “automobile” might be modeled as:

15

{automobile: *is a machine,*  
*is a vehicle,*  
*has engine,*  
*has brakes,*  
 ...  
 }

20

In principle, there might be infinitely many such properties, though practically speaking one might restrict the cardinality to  $\aleph_0$  (countably infinite) in order to ensure that the properties are addressable. If one were disposed to do so, one might require that there be only finitely many properties associated with a lexical element. However, there is no compelling reason to require finiteness.

25

At any rate, one can see that “automobile” is simply an element of the power set of  $P$ , the set of all propositions; i.e., it is an element of the set of all subsets of  $P$ . The power set is denoted as  $\wp(P)$ . Note that the first two properties of the “automobile” example express “*is a*” relationships. By “*is a*” is meant entailment. *Entailment* means that, were one to intersect the properties of every element of  $\wp(P)$  that is called, for example, “machine,” then the

30

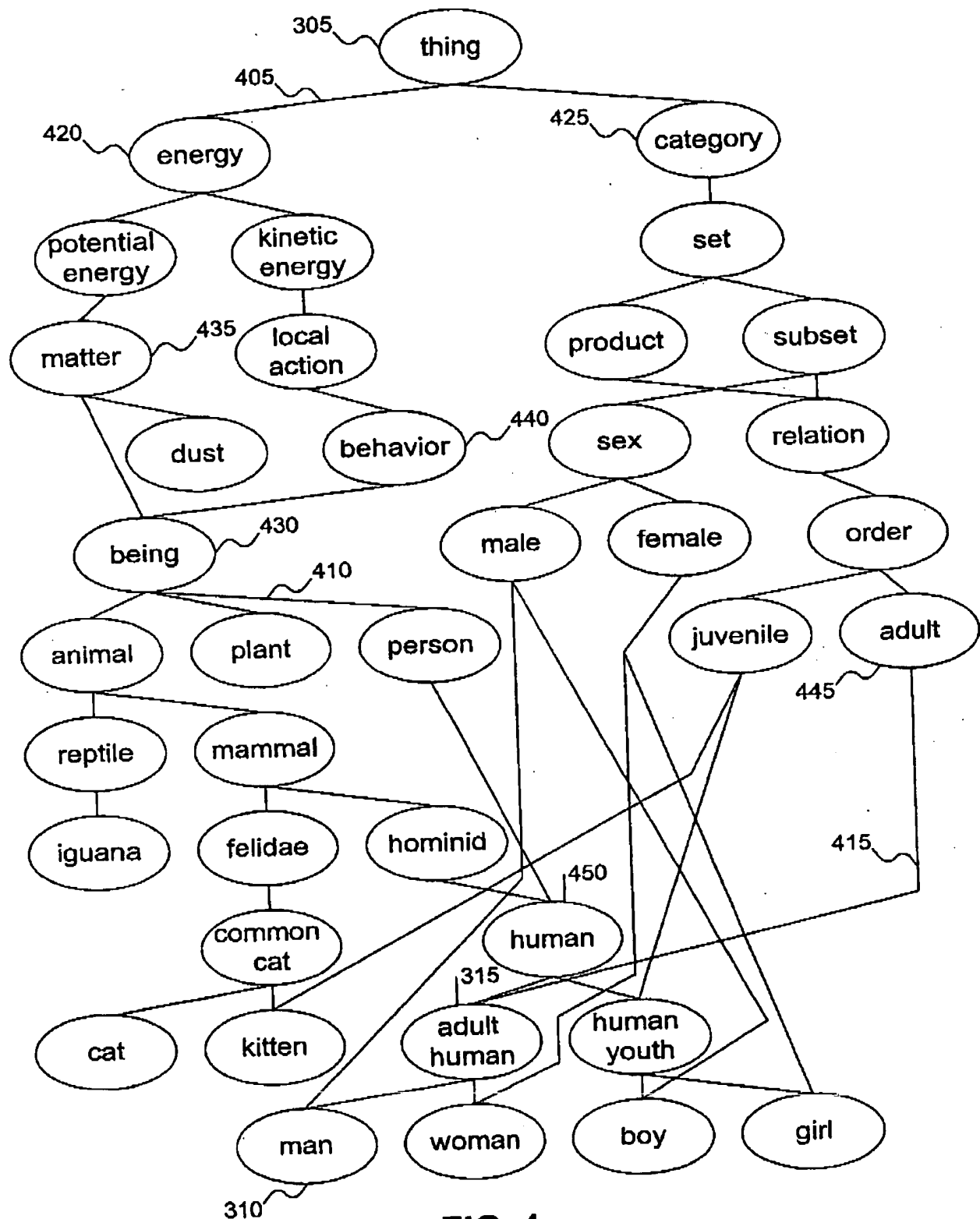


FIG. 4

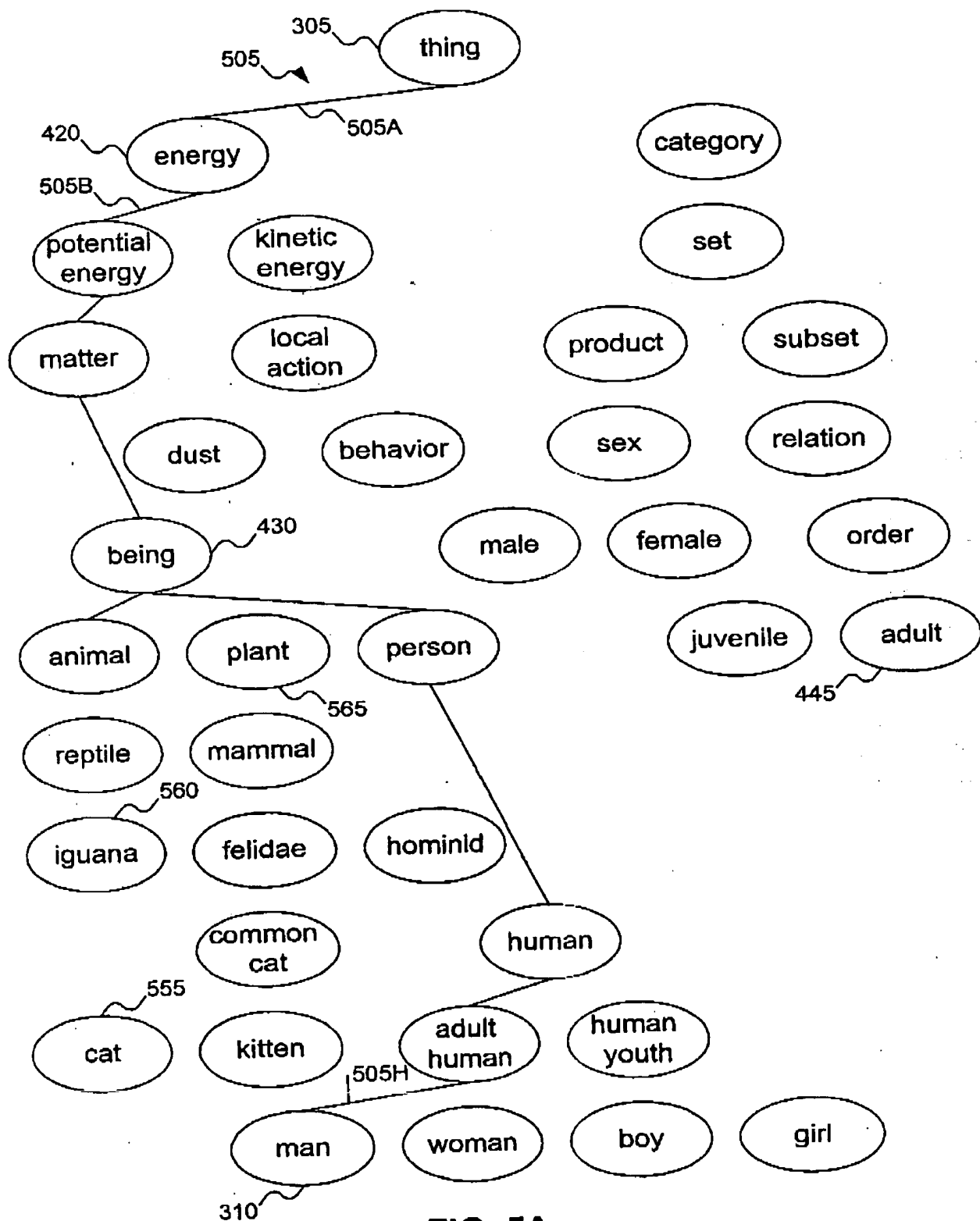


FIG. 5A



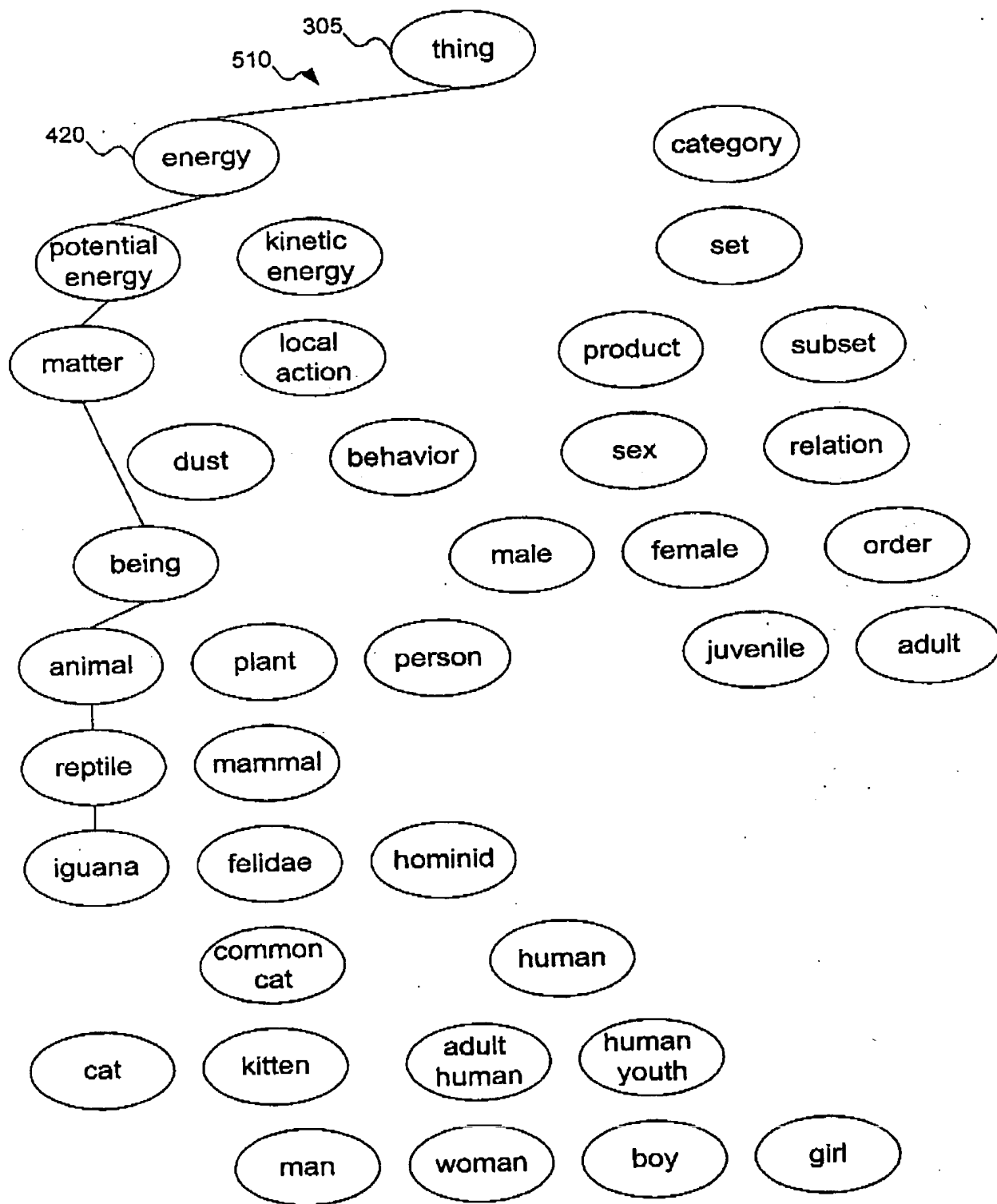


FIG. 5B

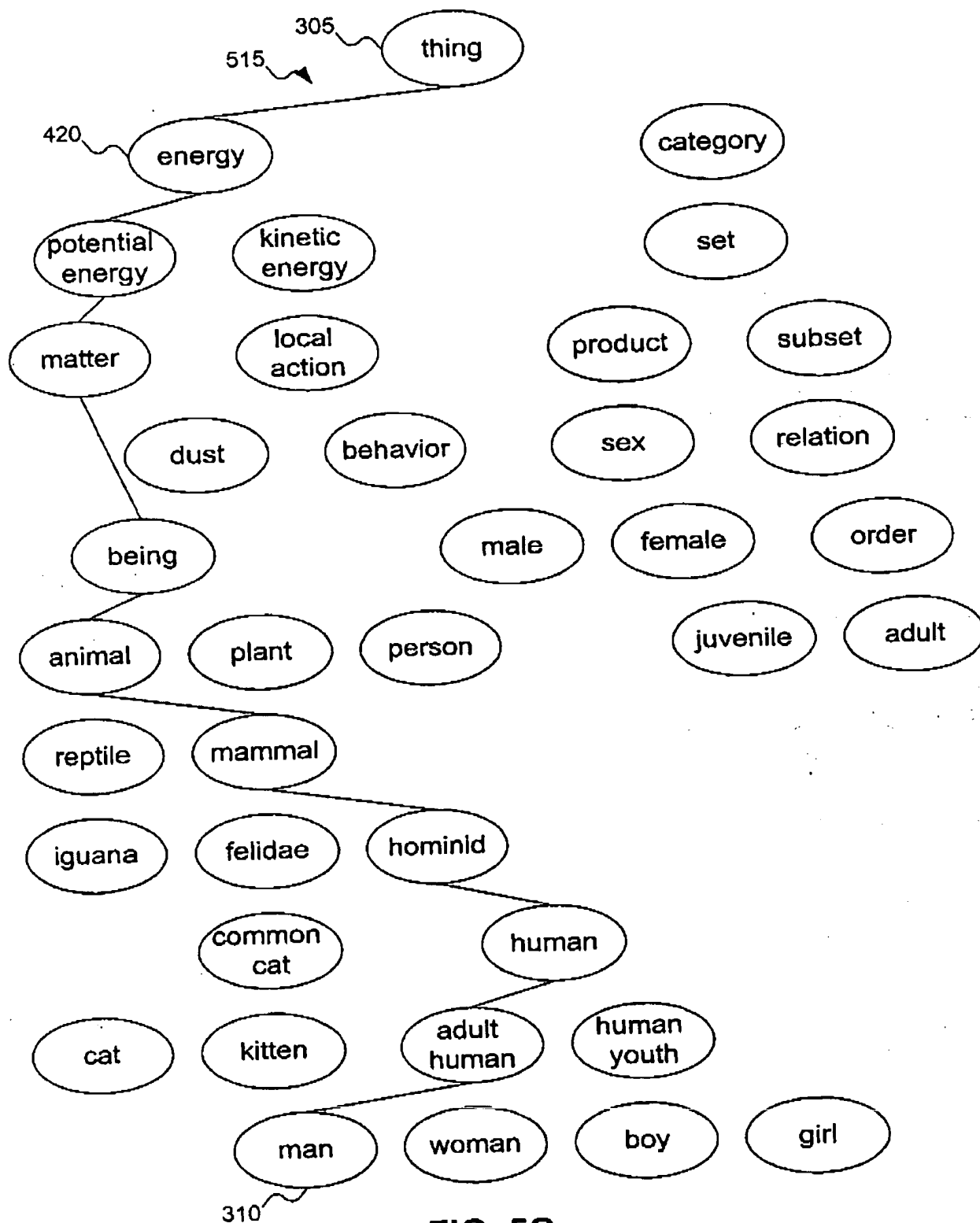


FIG. 5C

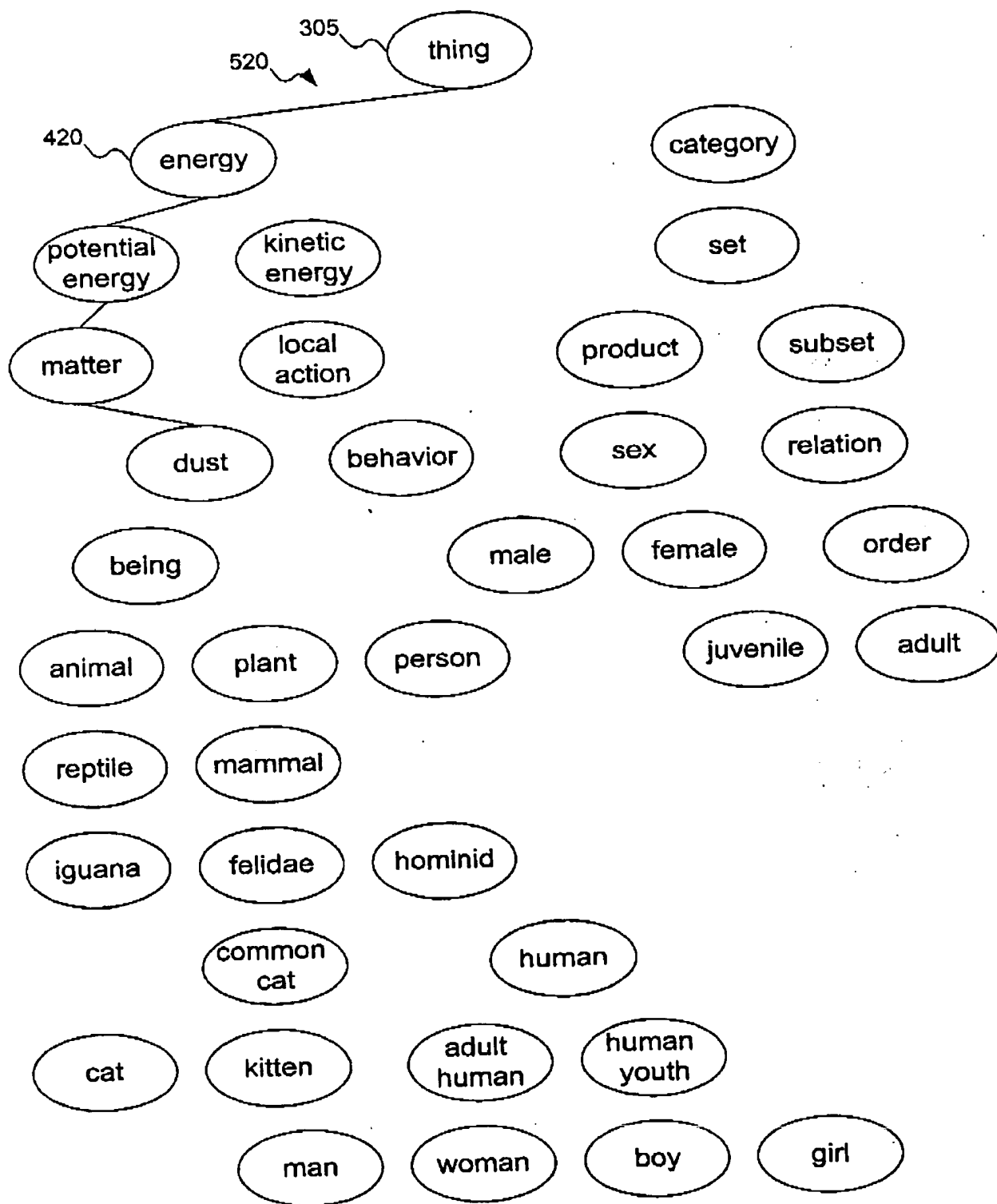


FIG. 5D

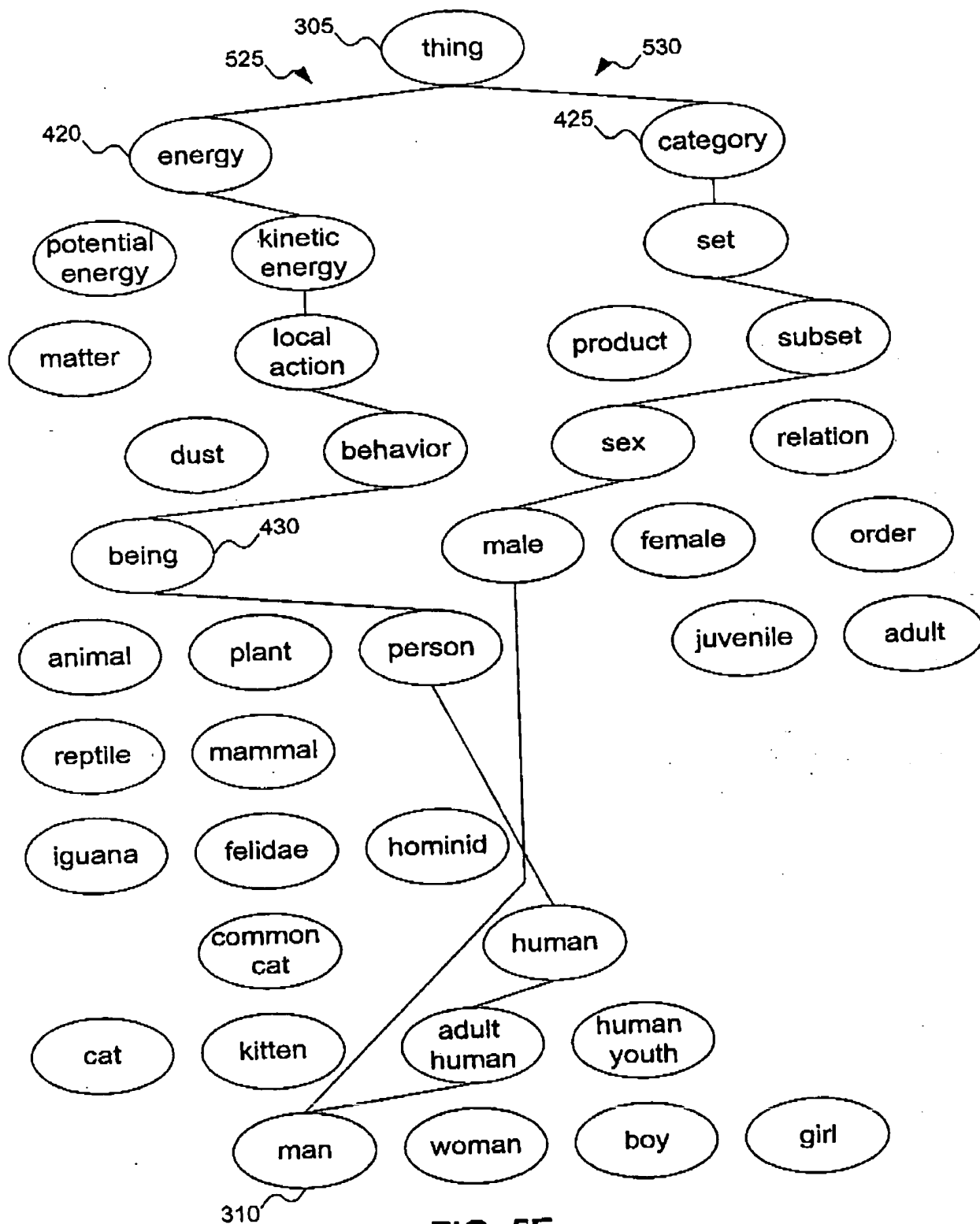
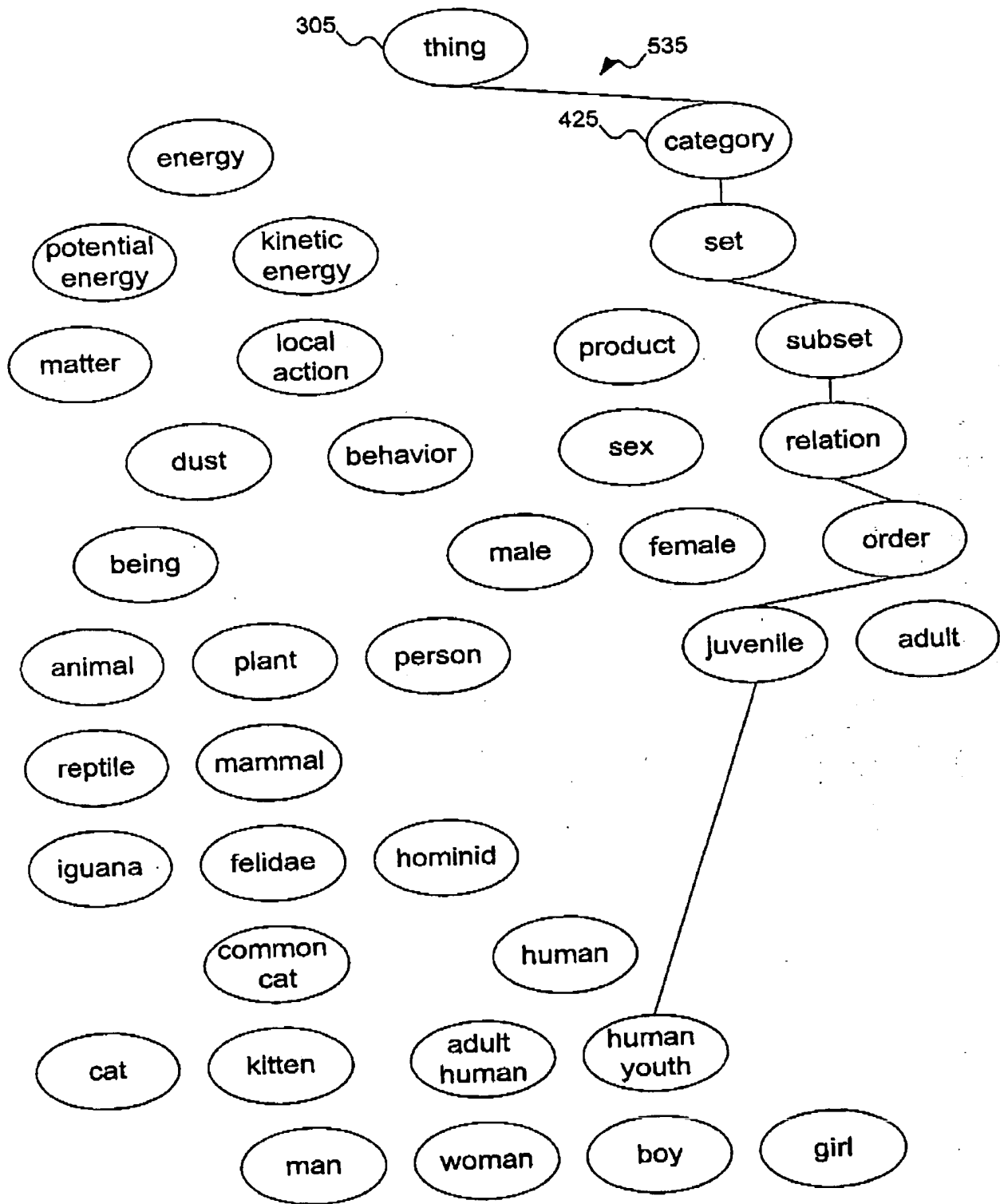


FIG. 5E



**FIG. 5F**

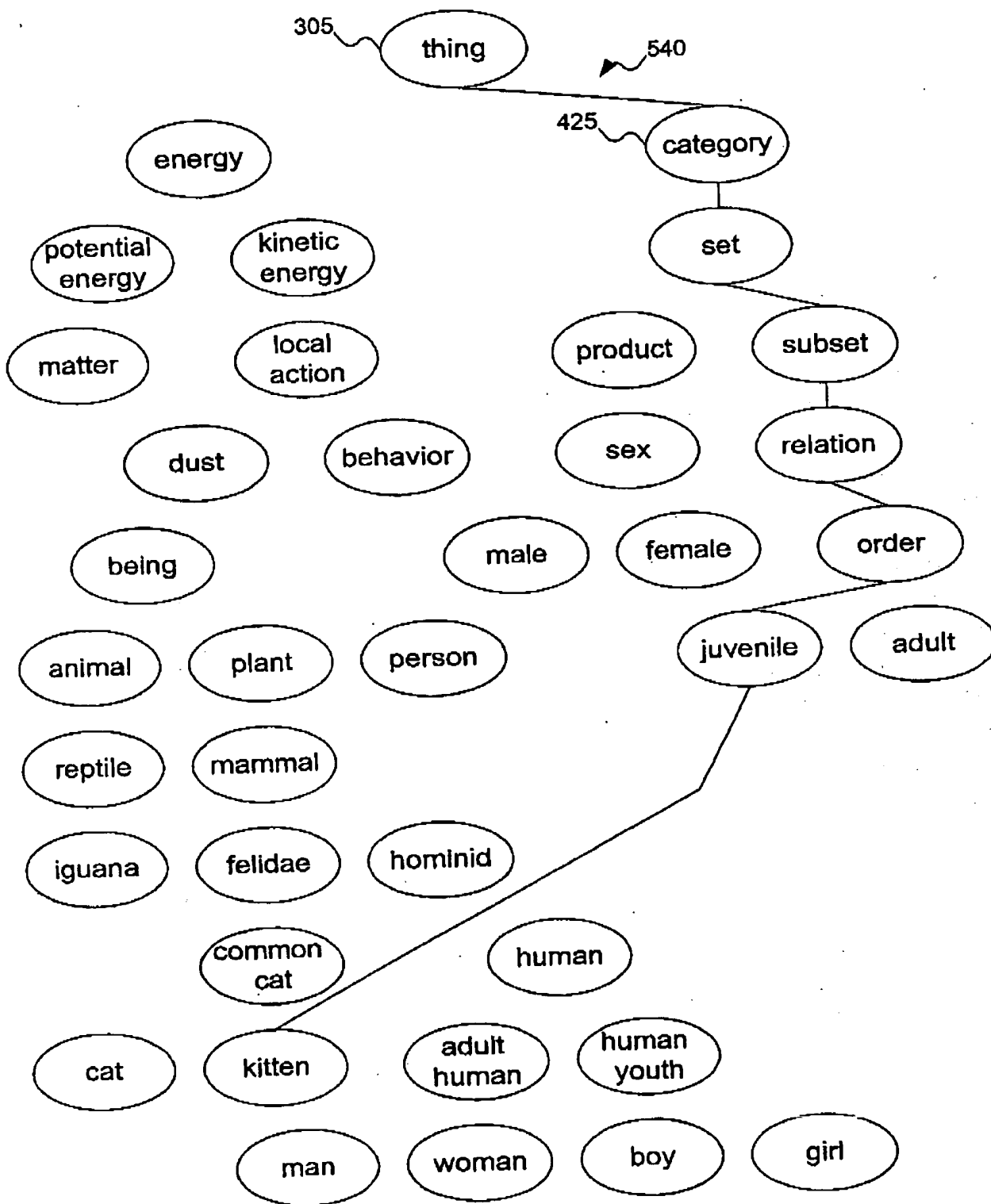


FIG. 5G